Learning Non-Parametric Basis Independent Models from Point Queries via Low-Rank Methods

LEARNING NON-PARAMETRIC BASIS INDEPENDENT MODELS FROM POINT QUERIES VIA LOW-RANK METHODS

HEMANT TYAGI AND VOLKAN CEVHER

ABSTRACT. We consider the problem of learning *multi-ridge* functions of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ from point evaluations of f. We assume that the function f is defined on an ℓ_2-ball in $\mathbb{R}^d$, g is twice continuously differentiable almost everywhere, and $\mathbf{A} \in \mathbb{R}^{k\times d}$ is a rank k matrix, where $k \ll d$. We propose a randomized, polynomial-complexity sampling scheme for estimating such functions. Our theoretical developments leverage recent techniques from low rank matrix recovery, which enables us to derive a polynomial time estimator of the function f along with uniform approximation guarantees. We prove that our scheme can also be applied for learning functions of the form: $f(\mathbf{x}) = \sum_{i=1}^{k} g_i(\mathbf{a}_i^T\mathbf{x})$, provided f satisfies certain smoothness conditions in a neighborhood around the origin. We also characterize the noise robustness of the scheme. Finally, we present numerical examples to illustrate the theoretical bounds in action.

recovery, randomized sampling, oracle-based learning

1. INTRODUCTION

Many important scientific and engineering problems revolve around models defined as multivariate continuous functions of d variables, where d is typically large. Examples include but are not limited to neural networks that are commonly used in pattern classification from data [1], path integrals with respect to Weiner measure that arise in the parameter estimation of stochastic processes [26], and smooth multivariate objective functions in optimization problems in machine learning and signal processing. As having an explicit form of a multivariate continuous function f alleviates analysis and computation in many applications, a great deal of research now focuses on learning such functions from their point values [9, 12, 29].

Unfortunately, even approximating multivariate continuous functions defined over classical unweighted spaces is in general intractable. This notion of intractability is precisely characterized by the *information complexity* of learning, which is defined as the minimum number of information extraction operations $n(e, d)$ that an algorithm performs to estimate a multivariate function within a uniform approximation error e [34]. If $n(e, d)$ depends exponentially on either e^{-1} or d, then the problem is called intractable. Polynomial tractability, on the other hand, specifically refers to the case when $n(e, d)$ depends polynomially on both d and e^{-1}. In the function learning setting, it is well known that the optimal order of the error of approximation for functions belonging to $\mathcal{C}^r[0, 1]^d$ is exponential: i.e.,

Key words and phrases. Multi-ridge functions, high dimensional function approximation, low rank matrix recovery, non linear approximation, oracle-based learning.

An extended abstract of this paper appeared in the 26$^{\text{th}}$ Annual Conference on Neural Information Processing Systems (NIPS), December 2012. The present draft is an expanded version with a more rigorous analysis and consists of proofs of all the results.

$n(e,d) = \Omega((1/e)^{d/r})$ for $e \in (0,1)$ (see [34] for example). As another example, [27] recently proved that the L_∞ approximation of $\mathcal{C}^\infty$ functions defined on $[0,1]^d$ is an intractable problem: i.e., $n(e,d) = \Omega(2^{\lfloor d/2 \rfloor})$ for $e \in (0,1)$. Therefore, further assumptions on the multivariate functions beyond smoothness are needed for the tractability of successful learning [15, 12, 9, 34].

Fortunately, many multivariate functions that arise in practice possess much more structure than an arbitrary d-variate continuous function. To this end, our work focuses on approximating a particular class of low dimensional functions known as *multi-ridge functions* with point queries. A multi-ridge function is a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$ defined using a $k \times d$, full rank matrix $\mathbf{A}$ as follows:

$$f(\mathbf{x}) \;=\; g(\mathbf{A}\mathbf{x}), \tag{1.1}$$

where g belongs to a restricted function class. Ridge functions are studied in Statistics under the name of "projection pursuit regression" [13, 11, 16]. The namesake was first introduced for the case $k = 1$ in 1975 by Logan and Shepp [23], in connection with the mathematics of computer tomography. Approximation theoretical questions regarding ridge functions have been studied in connection with the modeling of neural networks [28, 2], and also in ridgelets [4, 3]. A special case of (1.1) where f decomposes as:

$$f(\mathbf{x}) \;=\; \sum_{i=1}^{k} g_i(\mathbf{a}_i^T \mathbf{x}), \tag{1.2}$$

has several important applications in machine learning applications and are known as *multi-index* models in statistics and econometrics [41, 40, 20, 14].

Previous work. The recent literature can be split into two distinct camps with one taking an approximation theoretic view and the other pursuing a regression perspective.

In the approximation theoretic camp, the data is obtained with a sampling strategy tailored towards the structure of the underlying function f. [9] propose a greedy algorithm for estimating functions of the form $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$, where $g : [0,1] \to \mathbb{R}$ is a $\mathcal{C}^s$ function for $s \geq 1$. To establish tractable learning guarantees on f, the authors assume that $\mathbf{a}$ is stochastic, that is, $\mathbf{a} \succeq 0$ and $\mathbf{1}^T\mathbf{a} = 1$. They also assume $\mathbf{a}$ to be compressible, i.e., $\mathbf{a}$ lives in a weak ℓ_q-ball, and hence, can be well-approximated by a sparse set of its coefficients. In [12], the authors generalize the model of Cohen et al. to the matrix case (1.1) by assuming that each row of $\mathbf{A}$ is compressible without any sign restrictions and that g is in $\mathcal{C}^s$ for $s \geq 2$.

In the regression camp, the data is drawn independent and identically distributed (iid) from some unknown distribution. [29] leverage convex programming based on M-estimators, and study the sparse additive model, $f(\mathbf{x}) = \sum_{j \in S} g_j(x_j)$ ($|S| = k \ll d$), introduced by [22]. In this setting, [29] remove the smoothness assumptions on the function atoms g_j, and treat the case where g_j's lie in a reproducible Hilbert Kernel space. Moreover, [29] provide algorithm independent minimax approximation rates. For more examples in the regression camp, we refer the reader to [30, 17, 24, 22].

Our contributions. These works rigorously illustrate that it is highly advantageous to identify additional structures in the multivariate function for the tractability of learning. In this setting, our work belongs to the approximation theoretic camp and makes the following three contributions.

First, we generalize the approximation results of [12] to the class of $\mathcal{C}^2$ functions with arbitrary number of linear parameters k *without* the compressibility assumption on the rows of $\mathbf{A}$. To achieve this generalization, we leverage recent advances in the analysis of low-rank matrix recovery algorithms. As a result, we propose a stable, polynomial time algorithmic framework with a tractable sampling scheme, endowed with uniform approximation guarantees on f.

Second, we prove tractability of our framework for a wider function class - a key addition to the existing results which are limited to radial functions [12]. To acheive this we place second order conditions on f which are made clear in Proposition 2. As a side result, we are able to handle the important case of multi-index models (1.2). For instance, summation of k-kernel ridge functions (Epanechnikov, Gaussian, Cosine, etc.) functions are readily handled. This result also lifts the structure of sparse additive model from the regression camp to a basis free setting, but in turn restricts the functional atoms to be almost everywhere $\mathcal{C}^2$.

Third, we empirically illustrate the tightness of our sample complexity bounds on a variety of important function examples, such as logistic, quadratic forms, and summation of Gaussians. We also analytically show how additive white noise in the function queries impacts the sample complexity of our low-rank based approach.

Notation. We denote the ℓ_2-ball with radius $r > 0$ in $\mathbb{R}^d$ as $B_{\mathbb{R}^d}(r)$, and employ the shorthand $B_{\mathbb{R}^d}$ when $r = 1$. We use $\mu_{\mathbb{S}^{d-1}}$ for the uniform measure on the d-dimensional unit sphere $\mathbb{S}^{d-1}$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we let $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$ denote the inner product. We use $\ll \mathbf{X}, \mathbf{Y} \gg = \mathrm{Tr}(\mathbf{X}^T\mathbf{Y})$ as the standard matrix inner product where $\mathrm{Tr}(\cdot)$ is the matrix trace. $\|\mathbf{X}\|_*$ denotes the nuclear norm, $\|\mathbf{X}\|_F$ denotes the Frobenius norm, and $\|\mathbf{X}\|$ denotes the operator norm of $\mathbf{X}$. For any $\mathbf{x} \in \mathbb{R}^n$ we denote its ℓ_p norm by $\|\mathbf{x}\|_{\ell_p^n}$. For a given linear operator $\Phi : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, we use $[\Phi(\mathbf{X})]_i = \ll \Phi_i, \mathbf{X} \gg$ with $\Phi_i \in \mathbb{R}^{n_1 \times n_2}$, and denote $\Phi^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2}$ as the adjoint operator.

2. Setup and Assumptions

Problem statement. Broadly speaking, we are interested in deriving approximations for functions $f : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \to \mathbb{R}$ of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]^T$ is an arbitrary rank k matrix of dimensions $k \times d$. We restrict ourselves to the *oracle* setting where we can only extract information about f through its—possibly noisy—point evaluations.

Assumptions. We first assume $\mathbf{A}\mathbf{A}^T = \mathbf{I}_k$, where $\mathbf{I}$ is the $k \times k$ identity matrix. If this is not the case, we can express $\mathbf{A}$ through its singular value decomposition (SVD) as $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$ to obtain an equivalent representation: $f(\mathbf{x}) = g(\mathbf{U}\Sigma\mathbf{V}^T\mathbf{x}) = \bar{g}(\mathbf{V}^T\mathbf{x})$, where $\bar{g}(\mathbf{y}) = g(\mathbf{U}\Sigma\mathbf{y})$ and $\mathbf{y} \in B_{\mathbb{R}^k}(1 + \bar{\epsilon})$. It is straightforward to verify how our assumptions on g transfers on $\bar{g}$ (cf., [12]). While we discuss approximation results on A below, the readers should keep in mind that our final guarantees only apply to the function f and not necessarily for $\mathbf{A}$ and g individually.

We assume g to be a $\mathcal{C}^2$ function. By our set up, g also lives over a compact set, hence all its partial derivatives till the order of two are bounded as a result of the

Stone-Weierstrass theorem:

$$\sup_{|\beta|\leq 2} \left\|D^\beta g\right\|_\infty \leq C_2;\ D^\beta g \ = \ \frac{\partial^{|\beta|}}{\partial y_1^{\beta_1}\dots\partial y_k^{\beta_k}}\ ; \quad |\beta| = \beta_1 + \cdots + \beta_k$$

for some constant $C_2 > 0$. We also assume that an enlargement of the unit ball $B_{\mathbb{R}^d}$ on the domain of the function f for a sufficiently small $\bar{\epsilon} > 0$ is allowed. This is not a restriction, but is a consequence of our analysis as we work with directional derivatives of f at points on the unit sphere $\mathbb{S}^{d-1}$.

Our Ansatz. We verify the tractability of our sampling approach by checking whether or not the following Hessian matrix H is well-conditioned á la [12]:

$$H^f := \int_{\mathbb{S}^{d-1}} \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^T d\mu_{\mathbb{S}^{d-1}}(\mathbf{x}). \tag{2.1}$$

That is, for singular values of H^f, we have $\sigma_1(H^f) \geq \sigma_2(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha > 0$ for some α. We theoretically characterize the scaling of α in Section 5 for interesting classes of functions.

3. Oracle-based Low-Rank Learning of Multi-Ridge Functions

In this section, we first identify a first-order relationship in our learning problem that ties the function values at the point queries as an affine observation of a low-rank matrix, whose column space is equal to A^T. We then exploit this observation to motivate a class of polynomial time algorithms for approximate recovery of A. To establish algorithmic guarantees, we focus on a randomized sampling scheme that provides a bi-Lipschitz embedding of low rank matrices. We then provide an outline of our learning scheme, which we theoretically analyze in Section 4.

3.1. Observation and oracle models.

Our learning approach relies on a specific interaction of two sets: sampling centers and an associated set of directions for each center. Let us first denote the set of sampling centers as follows:

$$\mathcal{X} = \{\xi_j \in \mathbb{S}^{d-1}; j = 1, \dots, m_{\mathcal{X}}\}. \tag{3.1}$$

Along with each $\xi_j \in \mathcal{X}$, we define a directions matrix $\Phi_j = [\phi_{1,j}|\dots|\phi_{m_\Phi,j}]^T$, where $\phi \in B_{\mathbb{R}^d}(r)$ for some $r > 0$, which we specify in Section 3.3.

We now begin with a simple first order approximation of the function f as follows

$$f(\mathbf{x} + \epsilon\phi) = f(\mathbf{x}) + \epsilon \left\langle \phi, \nabla f(\mathbf{x}) \right\rangle + \epsilon E(\mathbf{x}, \epsilon, \phi), \tag{3.2}$$

where $\epsilon \ll 1$, and $\epsilon E(\mathbf{x}, \epsilon, \phi)$ is the approximation error. Substituting the ridge function form (1.1) into (3.2), we then stumble upon a perturbed observation model ($\nabla g(\cdot)$ is a $k \times 1$ vector) below

$$\left\langle \phi, A^T \nabla g(A\mathbf{x}) \right\rangle = \frac{1}{\epsilon} \left(f(\mathbf{x} + \epsilon\phi) - f(\mathbf{x}) \right) - E(\mathbf{x}, \epsilon, \phi). \tag{3.3}$$

Without loss of generality, we denote the evaluation of $f(\mathbf{x}+\epsilon\phi) - f(\mathbf{x})$ as a call to the oracle. When the oracle is flawless, then the error $E(\mathbf{x}, \epsilon, \phi)$ is characterized via Taylor's expansion:

$$E(\mathbf{x}, \epsilon, \phi) = \varepsilon := \frac{\epsilon}{2} \phi^T \nabla^2 f(\zeta(\mathbf{x}, \phi)) \phi, \tag{3.4}$$

where $\zeta(\xi,\phi) \in [\xi, \mathbf{x}+\epsilon\phi] \in B_{\mathbb{R}^d}(1+\epsilon r)$. In general, one can envision a noisy oracle providing imprecise function values. To address a broad set of cases, we modify the perturbation model as

$$E(\mathbf{x},\epsilon,\phi) = \varepsilon + \epsilon^{-1}\mathbf{z} + \mathbf{s}(\pi), \tag{3.5}$$

where $\mathbf{z} = \mathcal{N}(0,\sigma_z^2)$ is an iid, zero mean Gaussian noise with a variance parameter σ_z^2, and $\mathbf{s}$ is an unbounded *sparse* noise that either destroys the information in an oracle call with probability $\pi \ll 1$, or leaves it untouched with probability $1-\pi$. Section 4.4 further addresses the noise issues.

3.2. **Low-rank matrix recovery of A.** We now leverage (3.3) as a scaffold to derive our low-rank learning approach. We first introduce a rank-k matrix $\mathbf{X} := \mathbf{A}^T\mathbf{G}$ with $\mathbf{G} := [\nabla g(\mathbf{A}\xi_1)|\nabla g(\mathbf{A}\xi_2)|\cdots|\nabla g(\mathbf{A}\xi_{m_\mathcal{X}})]_{k\times m_\mathcal{X}}$. Based on (3.3), we then derive the following linear system of equations via the linear operator $\Phi : \mathbb{R}^{d\times m_\mathcal{X}} \to \mathbb{R}^{m_\Phi}$

$$\mathbf{y} = \Phi(\mathbf{X}) + E(\mathcal{X},\epsilon,\mathbf{\Phi}), \tag{3.6}$$

where we refer to $\mathbf{y} \in \mathbb{R}^{m_\Phi}$ as the (perturbed) measurements of $\mathbf{X}$.

The formulation (3.6) is known as the low-rank matrix recovery problem since the rank of the matrix $\mathbf{X}$ is $k \ll d$. In Appendix A, we explain three distinct low-rank recovery problem settings relevant to our problem, called affine rank minimization (ARM), matrix completion (MC), and robust principal component analysis (RPCA). Among these low-rank formulations, we focus on a randomized sampling scheme for the ARM problem using the matrix Dantzig selector for our derivations below. We leave the theoretical characterization the subset selection schemes for future.

3.3. **Low-rank matrix sampling.** It turns out that stable recovery of $\mathbf{X}$ from (3.6) is provable from number of measurements commensurate with the degrees of freedom in $\mathbf{X}$ (i.e., $m_\Phi = \mathcal{O}\left(k(d+m_\mathcal{X}-k)\right)$). By stable, we mean that the error of the estimated matrix in Frobenius norm is bounded by a constant times the Frobenius norm of the perturbations. Moreover, via the RPCA formulation, it is also possible to stably recover $\mathbf{X}$ even when a fraction of its entries are arbitrarily corrupted. These recovery guarantees of course are predicated upon the sampling scheme preserving the information in the low-rank matrix.

For concreteness, we require our sampling mechanism in this paper to provide a bi-Lipschitz embedding of all rank-r matrices $\mathbf{X}_r$ with overwhelming probability:

$$(1-\kappa_r)\left\|\mathbf{X}_r\right\|_F^2 \le \left\|\Phi(\mathbf{X}_r)\right\|_{l_2}^2 \le (1+\kappa_r)\left\|\mathbf{X}_r\right\|_F^2,$$

where κ_r is known as the the isometry constant [6]. We say that Φ satisfies the κ-RIP at rank r if $\kappa_r < \kappa$ where $\kappa \in (0,1)$. For the linear operator Φ to have κ-RIP, we form $\mathcal{X}$ by sampling points uniformly at random in $\mathbb{S}^{d-1}$ according to the uniform measure $\mu_{\mathbb{S}^{d-1}}$. We then construct the sampling directions for $i = 1,\ldots,m_\Phi,\ j = 1,\ldots,m_\mathcal{X},$ and $l = 1,\ldots,d$ as follows

$$\mathbf{\Phi} = \left\{\phi_{i,j} \in B_{\mathbb{R}^d}\left(\sqrt{d/m_\Phi}\right) : [\phi_{i,j}]_l = \pm\frac{1}{\sqrt{m_\Phi}} \text{with probability } 1/2\right\}. \tag{3.7}$$

As Φ is a Bernoulli random measurement ensemble it follows from standard concentration inequalities [31, 18] that for any rank-r $\mathbf{X} \in \mathbb{R}^{d\times m_\mathcal{X}}$

$$\mathbb{P}(\,|\,\left\|\Phi(\mathbf{X})\right\|_{\ell_2}^2 - \left\|\mathbf{X}\right\|_F^2\,| > t\left\|\mathbf{X}\right\|_F^2) \le 2e^{-\frac{m_\Phi}{2}(t^2/2-t^3/3)}, \quad t \in (0,1).$$

By using a standard covering argument as shown in Theorem 2.3 of [6] it is easily verifiable that Φ satisfies RIP with isometry constant $0 < \kappa_r < \kappa < 1$ with probability at least $1 - 2e^{-m_\Phi q(\kappa) + r(d+m_\mathcal{X}+1)u(\kappa)}$, where $q(\kappa) = \frac{1}{144}\left(\kappa^2 - \frac{\kappa^3}{9}\right)$ and $u(\kappa) = \log\left(\frac{36\sqrt{2}}{\kappa}\right)$.

3.4. **Our low-rank oracle learning scheme.** We outline the main steps involved in our approximation scheme in Algorithm 1. Step 1 is related to the sampling tractability of learning, which we study in Section 5. Step 2 forms the measurements based on the ARM formulation and our sampling scheme. Step 3 revolves around the ARM recovery, where we employ the matrix Dantzig selector algorithm for concreteness in our analysis. Step 4 maps the recovered low-rank matrix to A, followed by Step 5 that finally leads to the function estimate. Section 4 provides

Algorithm 1 Estimating $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$

1: Choose m_Φ and $m_\mathcal{X}$ (Section 5) and construct the sets $\mathcal{X}$ and $\mathbf{\Phi}$ (Section 3.3).
2: Choose ϵ (Section 4.2) and construct $\mathbf{y}$ using $y_i = \sum_{j=1}^{m_\mathcal{X}} \left[\frac{f(\xi_j + \epsilon\phi_{i,j}) - f(\xi_j)}{\epsilon}\right]$.
3: Obtain $\widehat{\mathbf{X}}$ via a stable low-rank recovery algorithm (Appendix A).
4: Compute SVD$(\widehat{\mathbf{X}}) = \widehat{\mathbf{U}}\widehat{\Sigma}\widehat{\mathbf{V}}^T$ and set $\widehat{\mathbf{A}}^T = \widehat{\mathbf{U}}^{(k)}$, corresponding to k largest singular values.
5: Obtain $\widehat{f}(\mathbf{x}) := \widehat{g}(\hat{\mathbf{A}}\mathbf{x})$ via quasi interpolants where $\widehat{g}(\mathbf{y}) := f(\widehat{\mathbf{A}}^T\mathbf{y})$.

an end-to-end analysis of the steps in Algorithm 1. Here, we further comment on two important ingredients in our learning scheme: the norm of the perturbations, and the function estimator in Step 5 of Algorithm 1 given an estimate $\widehat{A}$ of A.

Stability. We provide a stability characterization for the ARM recovery algorithms in the form of Proposition 1 below, which upperbounds the $\ell_2^{m_\Phi}$-norm of the noise ε for the perfect oracle setting.

Proposition 1. *In the factorization equality* (3.6)*, we have* $\|E\|_{\ell_2^{m_\Phi}} = \|\varepsilon\|_{\ell_2^{m_\Phi}} \leq \frac{C_2 \epsilon k^2}{2}\frac{m_\mathcal{X} d}{\sqrt{m_\Phi}}$.

Appendix B has the proof. Note that the dimension d appears in the bound as we do not make any compressibility assumption on $\mathbf{A}$. If the rows of $\mathbf{A}$ are compressible, that is $(\sum_{j=1}^d |a_{ij}|^q)^{1/q} \leq D_1 \ \forall\ i = 1, \ldots, k$ for some $0 < q < 1,\ D_1 > 0$, the bound becomes independent of d.

Our function estimator. Given $\widehat{\mathbf{A}}$ of $\mathbf{A}$ in Step 4, we construct $\widehat{f}(\mathbf{x}) := \widehat{g}(\widehat{\mathbf{A}}\mathbf{x})$ as our estimator, where $\widehat{g}(\mathbf{y}) := f(\widehat{\mathbf{A}}^T\mathbf{y})$ with $\mathbf{y} \in B_{\mathbb{R}^k}(1+\bar{\epsilon})$. We uniformly approximate the function $\widehat{g}$ by first sampling it on a rectangular grid : $h\mathbb{Z}^k \cap (-(1+\bar{\epsilon}), (1+\bar{\epsilon}))^k$ with uniformly spaced points in each direction (step size h). We then using quasi interpolants to interpolate in between the points thereby obtaining the approximation $\hat{g}_h$, where the complexity only depends on k. We refer the reader to Chapter 12 of [10] regarding the construction of these operators.

It is straightforward to prove that $\|\widehat{g} - \widehat{g}_h\|_\infty < Ch^2$, holds true for some constant C. By triangle inequality, we then carry the following approximation guarantee for $\widehat{g}_h$:

$$\|g - \widehat{g}_h\|_\infty \leq \|g - \widehat{g}\|_\infty + \|\widehat{g} - \widehat{g}_h\|_\infty .$$

In this loop, the samples of $\widehat{g}$ on the h-grid are obtained directly through point queries of f. However, the required number of samples for a given error depends only on k and not on d.

Remark 1. *(i) The parameter $\bar{\epsilon} = \epsilon\sqrt{d/m_\Phi}$, defining the domain of the function f, is bounded from above. In the course of deriving an approximation to f, we require ϵ to be at most $\mathcal{O}\left(\frac{1}{d}\sqrt{\frac{m_\Phi \alpha}{m_\mathcal{X}}}\right)$, (as is stated in Lemma 2), in order to obtain a non trivial approximation error guarantee. We shall also discover in Section 5 that α can be at most $\mathcal{O}(1)$ implying $\bar{\epsilon}$ to be typically at most $\mathcal{O}(1/\sqrt{d})$.*

(ii) As opposed to [12] *our scheme requires more number of sampling directions. To see this, observe that there is an underlying $d \times m_\mathcal{X}$ matrix $X = A^T G$ which contains information about the gradients of f at the sampled points $m_\mathcal{X}$. Here $\mathbf{G} := [\nabla g(\mathbf{A}\xi_1)|\nabla g(\mathbf{A}\xi_2)|\cdots|\nabla g(\mathbf{A}\xi_{m_\mathcal{X}})]_{k\times m_\mathcal{X}}$ and $\mathbf{A}$ is the underlying subspace matrix of size $k \times d$. Now in* [12]*, the compressibility assumption on the rows of $\mathbf{A}$ enables the authors to sample each column of X individually and then recover it using standard ℓ_1 minimisation. Note that each column of X is the linear combination of k-vectors each of which is compressible hence the resulting X will have compressible columns. In particular the same direction vector (generated at random) is used for measuring each column of X implying that for m_Φ measurements of the columns they need only m_Φ sampling directions. On the other hand we cannot do this since we make no compressibility assumption on $\mathbf{A}$. Hence we resort to taking linear measurements of the complete matrix X and aim to recover this matrix by employing low-rank matrix recovery algorithms. To obtain one measurement of X we need to generate $m_\mathcal{X}$ number of sampling directions implying that for m_Φ measurements of X we need $m_\mathcal{X} \times m_\Phi$ sampling directions.*

4. Analysis of Oracle-based Low-Rank Learning

In this section, the parameters involved our derivations are the dimension d of $\mathbf{x}$, the number of linear parameters k, the smoothness constant C_2 for the underlying function g, and the conditioning parameter $0 < \alpha < kC_2^2$ for H^f in (2.1). Section 5 unifies the results with our tractability claims.

4.1. Low-rank matrix recovery with Dantzig Selector.

In order to recover an approximation to the rank k matrix $\mathbf{X}$, we solve the nuclear norm minimization problem based on the following convex formulation [6]:

$$\widehat{\mathbf{X}}_{DS} = \arg\min \|M\|_* \text{ s.t. } \|\Phi^*\left(y - \Phi(M)\right)\| \leq \lambda, \tag{4.1}$$

where the optimal solution is the estimate $\widehat{\mathbf{X}}_{DS}$. This convex program is referred to as the *matrix Dantzig selector* [6]. While Appendix A lists a number of other convex formulations for low rank matrix recovery, we choose the matrix Dantzig selector for concreteness.

As in [6], we require the true matrix $\mathbf{X}$ to be feasible in the convex formulation, i.e., one should have $\|\Phi^*(\varepsilon)\| \leq \lambda$. In the case of bounded noise, Lemma 1 helps us choose this parameter whose proof is in Appendix C.

Lemma 1. *Given ε with a bounded $\ell_2^{m_\Phi}$ norm, it holds that $\|\Phi^*(\varepsilon)\| \leq \frac{C_2 \epsilon d m_\mathcal{X} k^2}{2\sqrt{m_\Phi}}(1+\kappa_1)^{1/2}$, with probability at least $1 - 2e^{-m_\Phi q(\kappa_1)+(d+m_\mathcal{X}+1)u(\kappa_1)}$.*

We now present the error bound for the matrix Dantzig selector as was obtained in [6] in Theorem 1. In Corollary 1, we exploit this result in our setting for $r = k$ in order to obtain the error bound for recovering the rank-k approximation $\widehat{\mathbf{X}}_{DS}^{(k)}$ to $\mathbf{X}$.

Theorem 1. *Let rank($\mathbf{X}$) $\leq r$ and let $\widehat{\mathbf{X}}_{DS}$ be the solution to (4.1). If $\kappa_{4r} < \kappa < \sqrt{2}-1$ and $\|\Phi^*(\varepsilon)\| \leq \lambda$, then we have with probability at least $1-2e^{-m_\Phi q(\kappa)+4r(d+m_\mathcal{X}+1)u(\kappa)}$ that*

$$\left\|\widehat{\mathbf{X}}_{DS} - \mathbf{X}\right\|_F^2 \leq C_0 r\lambda^2,$$

where C_0 depends only on the isometry constant κ_{4r}.

Corollary 1. *Denoting $\widehat{\mathbf{X}}_{DS}$ to be the solution of (4.1), if $\widehat{\mathbf{X}}_{DS}^{(k)}$ is the best rank-k approximation to $\widehat{\mathbf{X}}_{DS}$ in the sense of $\|\cdot\|_F$, and if $\kappa_{4k} < \kappa < \sqrt{2}-1$, then we have*

$$\left\|\mathbf{X} - \widehat{\mathbf{X}}_{DS}^{(k)}\right\|_F^2 \leq \frac{C_0 C_2^2 k^5 \epsilon^2 d^2 m_\mathcal{X}^2}{m_\Phi}(1+\kappa),$$

with probability at least $1 - 2e^{-m_\Phi q(\kappa)+4k(d+m_\mathcal{X}+1)u(\kappa)}$, where the constant C_0 depends only on κ_{4k}.

Corollary 1 is the main result of this subsection, which is proved in Appendix D.

4.2. **Approximation of A.** In the previous subsection, we derive a rank-k approximation $\widehat{\mathbf{X}}_{DS}^{(k)}$ of the original rank-k matrix $\mathbf{X}$ with a bound on the approximation error $\left\|\widehat{\mathbf{X}}_{DS}^{(k)} - \mathbf{X}\right\|_F$. Here, we are interested in recovering an approximation $\widehat{\mathbf{A}}$ to the matrix $\mathbf{A}$ from $\widehat{\mathbf{X}}_{DS}^{(k)}$. Trivially, this can be achieved by setting $\widehat{\mathbf{A}}$ to the left singular vector matrix of $\widehat{\mathbf{X}}_{DS}^{(k)}$. The purpose of the analysis here is to theoretically characterize the ensuing approximation error.

Let the SVD of $\mathbf{X}$ and $\widehat{\mathbf{X}}_{DS}^{(k)}$ be $\mathbf{X} = \mathbf{A}^T\mathbf{G} = \mathbf{A}^T\mathbf{U}_G\Sigma_G\mathbf{V}_G^T = \mathbf{A}_1^T\Sigma_G\mathbf{V}_G^T$ and $\widehat{\mathbf{X}}_{DS}^{(k)} = \widehat{\mathbf{A}}^T\widehat{\Sigma}\widehat{\mathbf{V}}$, respectively. Then, $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ and $\widehat{\Sigma} = \text{diag}(\widehat{\sigma}_1, \widehat{\sigma}_2, \ldots, \widehat{\sigma}_k)$ are diagonal matrices with $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_k$ and $\widehat{\sigma}_1 \geq \widehat{\sigma}_2 \geq \ldots \widehat{\sigma}_k$, respectively. Moreover, $\mathbf{U}_G$ is a $k \times k$ unitary matrix. The columns of $\mathbf{A}_1^T, \mathbf{V}_G^T$ and $\widehat{\mathbf{A}}, \widehat{\mathbf{V}}$ are the singular vectors of $\mathbf{X}$ and $\widehat{\mathbf{X}}_{DS}^{(k)}$, respectively. Finally, we have $\sigma_i = \sqrt{\lambda_i(\mathbf{G}\mathbf{G}^T)}$ where λ_i denotes the i^{th} eigenvalue of

$$\mathbf{G}\mathbf{G}^T = \sum_{j=1}^{m_\mathcal{X}} \left(\nabla g(\mathbf{A}\xi_j)\nabla g(\mathbf{A}\xi_j)^T\right). \tag{4.2}$$

We now show that if $\left\|\mathbf{X} - \widehat{\mathbf{X}}_{DS}^{(k)}\right\|_F$ is driven to be smaller than a threshold then it leads to a probabilistic lower bound on $\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F$. Lemma 2, proved in Appendix E, precisely states this fact.

Lemma 2. *For a fixed $0 < \rho < 1$, $m_\mathcal{X} \geq 1$, $m_\Phi < m_\mathcal{X} d$ if $\epsilon < \frac{1}{C_2 k^2 d(\sqrt{k}+\sqrt{2})}\left(\frac{(1-\rho)m_\Phi \alpha}{(1+\kappa)C_0 m_\mathcal{X}}\right)^{1/2}$, then with probability at least $1-k\exp\left\{-\frac{m_\mathcal{X}\alpha\rho^2}{2kC_2^2}\right\}-2\exp\left\{-m_\Phi q(\kappa)+4k(d+m_\mathcal{X}+1)u(\kappa)\right\}$*

we have

$$\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F \geq \left(k - \frac{2\tau^2}{(\sqrt{(1-\rho)m_{\mathcal{X}}\alpha} - \tau)^2}\right)^{1/2},$$

where $\tau^2 = \frac{C_0 C_2^2 k^5 \epsilon^2 d^2 m_{\mathcal{X}}^2}{m_\Phi}(1+\kappa)$ *is the error bound derived in Corollary 1.*

Choice of ϵ. We note here that a guaranteed lower bound on $\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F$, of say $(k\eta)^{1/2}$ for some $0 < \eta < 1$, follows along the lines of the proof in Appendix E by ensuring that the following holds:

$$\epsilon < \frac{1}{C_2 k^2 d(\sqrt{k(1-\eta)} + \sqrt{2})}\left(\frac{(1-\rho)m_\Phi \alpha(1-\eta)}{(1+\kappa)C_0 m_{\mathcal{X}}}\right)^{1/2}.$$

4.3. **Approximation of f.** We now have the necessary background to state our main approximation result for the function f.

Theorem 2. *(Main approximation theorem) Let us fix* $\delta \in \mathbb{R}^+$, $0 < \rho < 1, 0 < \kappa < \sqrt{2}-1$. *Under the assumptions and notations mentioned earlier, for a fixed* $m_{\mathcal{X}} \geq 1$, $m_\Phi < m_{\mathcal{X}} d$ *and* $\epsilon < \frac{\delta}{C_2 k^{5/2} d(\delta + 2C_2\sqrt{2k})}\left(\frac{(1-\rho)m_\Phi \alpha}{(1+\kappa)C_0 m_{\mathcal{X}}}\right)^{1/2}$ *we have that the function* $\widehat{f}(\mathbf{x}) = \widehat{g}(\widehat{\mathbf{A}}\mathbf{x})$ *defined by means of* $\widehat{g}(y) := f(\widehat{\mathbf{A}}^T y), \quad \mathbf{y} \in B_{\mathbb{R}^k}(1+\bar{\epsilon})$ *has the uniform approximation bound*

$$\left\|f - \widehat{f}\right\|_\infty \leq \delta,$$

with probability at least $1 - k\exp\left\{-\frac{m_{\mathcal{X}}\alpha\rho^2}{2kC_2^2}\right\} - 2\exp\left\{-m_\Phi q(\kappa) + 4k(d + m_{\mathcal{X}} + 1)u(\kappa)\right\}$.

We provide the proof of our main approximation result Theorem 2 in Appendix F. In Section 5, we establish the tractability of our learning algorithm and also provide a comparison of our sampling bounds with those of [12] (i.e. $\mathbf{A}$ is compressible) for different function classes. In particular, we show that our sampling bounds can be better than [12] depending on the compressibility of $\mathbf{A}$. For instance, if $1 < q < 2$, then our bounds exhibit better scaling. Furthermore the results of [12] also benefit from our proposition that shows how the parameter α behaves for a variety of models such as the class of additive function models.

Remark 2. *(i) We can also consider approximating functions of the form:* $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$, *assuming without loss of generality that* $\|\mathbf{b}\|_{\ell_2^k} \leq 1$. *Then, our estimator* $\widehat{f}$ *attains the following form:* $\widehat{f}(\mathbf{x}) = \widehat{g}(\widehat{\mathbf{A}}\mathbf{x}) = g(\mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\mathbf{x} + \mathbf{b})$, *where* $\widehat{g}(y) := f(\widehat{\mathbf{A}}^T y),\ \mathbf{y} \in B_{\mathbb{R}^k}(1+\bar{\epsilon})$. *It is straightforward to verify that we obtain the same approximation bound on* $\left\|f - \widehat{f}\right\|_\infty$ *along the lines of the proof of Theorem 2. Furthermore, we can then uniformly approximate the function* $\hat{g}$ *by first sampling it on a rectangular grid* $h\mathbb{Z}^k \cap (-(2+\bar{\epsilon}), (2+\bar{\epsilon}))^k$ *as before with uniformly spaced points in each direction. Subsequently, by using quasi interpolants to interpolate between the points we obtain an approximation* $\hat{g}_h$. *In this particular setting, we need not approximate* $\mathbf{b}$ *to derive approximation guarantees on* f. *In particular we need only use a bound on* $\|\mathbf{b}\|_{\ell_2^k}$ *to accordingly set the size of the sampling grid.*

(ii) In Theorem 2 the step size parameter ϵ needs to be suitably small in order to guarantee the approximation result on f. This suugests that for large d, the requirement on ϵ might be too strict leading to numerical issues in approximating the gradient of f by finite differences as in (3.3)*. However note that the bound on ϵ depends on the ratio $\sqrt{m_\Phi/m_\mathcal{X}}$. Hence one can also choose a constant ϵ and $m_\mathcal{X} = O(1/\alpha)$. We can then compensate the choice of ϵ by choosing a suitably large value of m_Φ (as determined from Theorem 2 by the parameters $k, d, C_2, C_0, \rho, \kappa, \delta$ and α) resulting in a good approximation to f with high probability. We also note in our numerical simulations in Section 6 that it suffices to consider reasonable values such as $\epsilon \sim 10^{-3}$ which leads to stable approximation results.*

4.4. **Impact of measurement noise on learning scheme.** For the simplicity of our subsequent theoretical analysis, we fix ϵ as a small constant. As a by-product, ϵ^{-1} linearly amplifies the oracle Gaussian noise within the perturbation model (3.5). This is inherently due to the way we leverage the oracle calls while forming our naive gradient estimates: $\epsilon^{-1}\left(f(\mathbf{x}+\epsilon\phi) - f(\mathbf{x})\right)$. We note, however, that there are much better ways in practice to exploit the noisy oracle values to obtain de-noised gradient estimates by adaptively varying the region size and collectively using the oracle values (e.g., in the manner of regression methods in statistics or trust-region methods in optimization). Of course, the ideal solution in our formulation is to have access to a *gradient oracle*, which has small perturbations. We now further address these issues here.

Gaussian noise. Let us first assume that the evaluation of f at a point $\mathbf{x} \in B_{\mathbb{R}^d}(1+\bar{\epsilon})$ yields: $f(\mathbf{x})+Z$, where $Z \sim \mathcal{N}(0,\sigma^2)$. Thus under this noise model, (3.6) changes to:

$$\Phi(\mathbf{X}) = \mathbf{y} + \varepsilon + \mathbf{z} \tag{4.3}$$

where $\mathbf{z} \in \mathbb{R}^{m_\Phi}$ and $z_i = \sum_{j=1}^{m_\mathcal{X}} \frac{z_{ij}}{\epsilon}$. Assuming the iid noise samples, we have $z_{ij} \sim \mathcal{N}(0, 2\sigma^2)$, and $z_i \sim \mathcal{N}\left(0, \frac{2m_\mathcal{X}\sigma^2}{\epsilon^2}\right)$ for $i = 1, \ldots, m_\Phi$. Therefore, the noise variance gets amplified by a polynomial factor $\frac{m_\mathcal{X}}{\epsilon^2}$.

In our analysis, the parameter ϵ is assumed to be sufficiently small. In fact, Lemma 2 requires

$$\epsilon < \frac{1}{C_2 k^2 d(\sqrt{k}+\sqrt{2})}\left(\frac{(1-\rho)m_\Phi \alpha}{(1+\kappa)C_0 m_\mathcal{X}}\right)^{1/2}.$$

Therefore, for large d, ϵ is at most $\mathcal{O}\left(\frac{\alpha^{1/2}}{d}\right)$. To make the matters worse, the next section shows that α can be at most $\mathcal{O}(1)$ and usually decays polynomially with d. Thus, we see that the noise variance gets amplified as the dimension d and the number of samples $m_\mathcal{X}$ increases.

To further elaborate on how this affects the low rank recovery scheme, recall that in the convex program (4.1), we require the true matrix $\mathbf{X}$ to be feasible. In the setting of (4.3), this behooves us to consider $\|\Phi^*(\varepsilon + \mathbf{z})\| \leq \lambda$ for the feasibility of the solution. Let $m = \max(m_\Phi, m_\mathcal{X})$. Then, Lemma 1.1 [6] leads to the following bound with high probability ($\gamma > 2\sqrt{\log 12}$)

$$\|\Phi^*(\mathbf{z})\| \leq 2\gamma\sqrt{(1+\kappa_1)m}\sqrt{\frac{2m_\mathcal{X}\sigma^2}{\epsilon^2}}$$

Using this with result of Lemma 1, the following bound holds with high probability for $\gamma > 2\sqrt{\log 12}$

$$\|\Phi^*(\varepsilon + \mathbf{z})\| \leq \frac{2\gamma\sigma}{\epsilon}\sqrt{2m(1+\kappa_1)m_{\mathcal{X}}} + \frac{C_2\epsilon d m_{\mathcal{X}} k^2}{2\sqrt{m_\Phi}}(1+\kappa_1)^{1/2}.$$

We observe that as opposed to the perfect oracle setting we can no longer control the upper bound on $\|\Phi^*(\varepsilon + \mathbf{z})\|$ by simply reducing ϵ, due to the appearance of the $(1/\epsilon)$ term. Hence, unless σ is $\mathcal{O}(\epsilon)$ or less, (e.g., σ reduces with d), we can declare that our learning scheme with the matrix Dantzig selector is sensitive to noise, also when we use the minimum number of samples for recovery and we do not change the way we calculate the gradients. However, in many practical cases, it is possible to increase the number samples by a factor of d since noisy oracles tend to be cheaper. Alternatively, we must leverage the noisy oracle samples with more sophisticated methods to obtain denoised gradient estimates. Hence, for additional stability against Gaussian oracles with a constant noise variance, our tractability results in Section 5 needs to multiplied by a polynomial factor of d.

5. Information Complexity of Oracle-based Low-Rank Learning

In this section, we establish the tractability of our approximation strategy. As the first step, we note that the uniform approximation result in Theorem 2 holds with probability $1 - p_1 - p_2$ when

$$m_{\mathcal{X}} > \frac{2kC_2^2}{\alpha\rho^2}\log(k/p_1), \quad m_\Phi > \frac{\log(2/p_2) + 4k(d + m_{\mathcal{X}} + 1)u(\kappa)}{q(\kappa)}. \tag{5.1}$$

Therefore, for a desired probability of success, the sampling complexities scales as $m_{\mathcal{X}} = \mathcal{O}\left(\frac{k\log k}{\alpha}\right)$ and $m_\Phi = \mathcal{O}(k(d + m_{\mathcal{X}}))$ for large d. At this juncture, while we seemingly have the complexity of our randomized sampling scheme in Section 3.3, the effect of the parameter α is still implicit.

Appendix G relates the parameter α to the Hessian matrix H^f in our Ansatz in Section 2. Based on this discussion, we can rigorously observe that the conditioning of the matrix H^f for large d would be determined predominantly by the behavior of g in a open neighborhood around the origin. This behavior is quite straightforward to analyze when $k = 1$. What is not so easy to characterize is the behavior when $k > 1$. For instance, [12] finds it necessary to further constrain f to be a radial function to analyze the behavior of α when $k > 1$. By radial function, we mean $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}) = g_0(\|\mathbf{A}\mathbf{x}\|_{l_2^k})$, where g_0 is $\mathcal{C}^2$ smooth due to our problem set up.

One of the main contributions in this work is that we provide a local condition in Proposition 2 below (proved in Appendix H) that alleviates required conditions on the global structure of f:

Proposition 2. *Assume that $g \in \mathcal{C}^2 : B_{\mathbb{R}^k} \to \mathbb{R}$ has Lipschitz continuous second order partial derivatives in an open neighborhood of the origin, $\mathcal{U}_\theta = B_{\mathbb{R}^k}(\theta)$ for some fixed θ (depending only on k with k fixed):*

$$\frac{\left|\frac{\partial^2 g}{\partial y_i \partial y_j}(\mathbf{y}_1) - \frac{\partial^2 g}{\partial y_i \partial y_j}(\mathbf{y}_2)\right|}{\|\mathbf{y}_1 - \mathbf{y}_2\|_{l_2^k}} < L_{i,j} \quad \forall \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{U}_\theta, \mathbf{y}_1 \neq \mathbf{y}_2,\ i,j = 1,\ldots,k.$$

Denoting $L = \max_{1\leq i,j\leq k} L_{i,j}$, assume that $\nabla^2 g(\mathbf{0})$ is full rank, and either one of the following conditions hold:

(1) $\nabla g(\mathbf{0}) = \mathbf{0}$.
(2) $\nabla g(\mathbf{0}) \neq \mathbf{0}$ *and* $L = O(1/d)$.

Then, we have $\alpha = \Theta(1/d)$ *as* $d \to \infty$.

We are now ready to consider example function classes for $k = 1$ as well as $k > 1$ below, and derive the sampling complexities. As a baseline, we compare each result with [12] to highlight the variations as a result of forgoing the compressibility assumption on A.

5.1. **Function classes for $k = 1$.** [12] defines the following sets of classes of $\mathcal{C}^2$ smooth ridge functions for the case $k = 1$, for which they establish the scaling behavior of α to be polynomial in $1/d$:

(1) $[0 < q < 1,\ C_1 > 1$ and $C_2 \geq \alpha_0 > 0]$: $\mathcal{F}_d^1 := \mathcal{F}_d^1(\alpha_0, q, C_1, C_2) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} | \exists \mathbf{a} \in \mathbb{R}^d, \|\mathbf{a}\|_{\ell_2^d} = 1, \|\mathbf{a}\|_{\ell_q^d} \leq C_1$ and $\exists g \in \mathcal{C}^2(B_{\mathbb{R}}), |g'(\mathbf{0})| \geq \alpha_0 > 0 : f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})\}$.
(2) [For an open neighborhood $\mathcal{U}$ of 0, $0 < q < 1$, $C_1 > 1$, $C_2 \geq \alpha_0 > 0$ and $M \in \mathbb{N}$]: $\mathcal{F}_d^2 := \mathcal{F}_d^2(\mathcal{U}, \alpha_0, q, C_1, C_2, M) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} : \exists \mathbf{a} \in \mathbb{R}^d, \|\mathbf{a}\|_{l_2^d} = 1, \|\mathbf{a}\|_{l_q^d} \leq C_1$ and $\exists g \in \mathcal{C}^2(B_{\mathbb{R}}) \bigcap \mathcal{C}^{M+2}(\mathcal{U}), g^{(N)}(\mathbf{0}) = 0 \quad \forall \quad 1 \leq N \leq M, \left|g^{(M+1)}(\mathbf{0})\right| \geq \alpha_0 > 0 : f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})\}$.

We now generalize the above two classes in two non-trival ways:

(1) By doing away with the compressibility assumption on $\mathbf{a}$ from both $\mathcal{F}_d^1$ and $\mathcal{F}_d^2$.
(2) By showing along the lines of the proof of Proposition 2 that in $\mathcal{F}_d^2$, one can relax the space: $\mathcal{C}^2(B_{\mathbb{R}}) \bigcap \mathcal{C}^{M+2}(\mathcal{U})$ to $\mathcal{C}^2(B_{\mathbb{R}}) \bigcap \mathcal{C}^{M+1}(\mathcal{U}) \bigcap \mathcal{L}^{M+1}(\mathcal{U}, L)$. Here $\mathcal{L}^{M+1}(\mathcal{U}, L)$ denotes the space of $\mathcal{C}^{M+1}(\mathcal{U})$ functions whose $(M+1)^{th}$ derivatives are Lipschitz continuous with constant L.

For the sake of completeness, here are our generalized function classes:

(1) $[C_2 \geq \alpha_0 > 0]$: $\mathcal{H}_d^1 := \mathcal{H}_d^1(\alpha_0, C_2) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} | \exists \mathbf{a} \in \mathbb{R}^d, \|\mathbf{a}\|_{l_2^d} = 1,$ and $\exists g \in \mathcal{C}^2(B_{\mathbb{R}}), |g'(\mathbf{0})| \geq \alpha_0 > 0 : f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})\}$.
(2) [For an open neighborhood $\mathcal{U}$ of 0, $C_2 \geq \alpha_0 > 0$, $0 < L < \infty$ and $M \in \mathbb{N}$]: $\mathcal{H}_d^2 := \mathcal{H}_d^2(\mathcal{U}, \alpha_0, C_2, M, L) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} : \exists \mathbf{a} \in \mathbb{R}^d, \|\mathbf{a}\|_{l_2^d} = 1$ and $\exists g \in \mathcal{C}^2(B_{\mathbb{R}}) \bigcap \mathcal{C}^{M+1}(\mathcal{U}) \bigcap \mathcal{L}^{M+1}(\mathcal{U}, L), g^{(N)}(\mathbf{0}) = 0 \quad$ for all $\quad 1 \leq N \leq M, \left|g^{(M+1)}(\mathbf{0})\right| \geq \alpha_0 > 0 : f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})\}$.

Table 1 summarizes the sampling complexities for the above function classes. Observe that the sampling complexity increases from $\mathcal{O}(\log d)$ to $\mathcal{O}(d)$ when $g'(0) \neq 0$ and from $\mathcal{O}(d^{\frac{2M}{2-q}})$ to $\mathcal{O}(d^{2M})$ when the first M order partial derivatives of g at the origin are 0.

5.2. **Function classes for $k > 1$.** The case $k > 1$ is significantly more challenging to handle as compared to the case $k = 1$. [12] shows that if f is a radial function, $f(\mathbf{x}) = g(\mathbf{Ax}) = g_0(\|\mathbf{Ax}\|_{l_2^k})$, where g_0 is $\mathcal{C}^2$, then they can handle the following scenario depending on the local smoothness properties of g_0:

[For an open neighborhood $\mathcal{U}$ of 0]: $\mathcal{G}_{d,k} := \{M \in \mathbb{N}, g_0 \in \mathcal{C}^2(B_{\mathbb{R}}) \bigcap \mathcal{C}^{M+2}(\mathcal{U}), g_0^{(N)}(0) = 0 \quad \forall 1 \leq N \leq M$ and $\left|g_0^{(M+1)}(0)\right| \geq \alpha_0 > 0\}$.

In particular the authors show that for the above function class, $\alpha = \Theta(d^{-M})$. The proof of this result can be found in Section 4.3 of [12]. Table 2 provides a

Function class	Scaling of α	$m_{\mathcal{X}}$	m_Φ	$m_{\mathcal{X}} \times (m_\Phi + 1)$
$\mathcal{F}_d^1$	$\Theta(1)$	$\mathcal{O}(1)$	$\mathcal{O}(\log d)$	$\mathcal{O}(\log d)$
$\mathcal{H}_d^1$	$\Theta(1)$	$\mathcal{O}(1)$	$\mathcal{O}(d)$	$\mathcal{O}(d)$
$\mathcal{F}_d^2$	$\Theta(d^{-M})$	$\mathcal{O}(d^M)$	$\mathcal{O}\left(d^{\frac{Mq}{2-q}}\right)$	$\mathcal{O}\left(d^{\frac{2M}{2-q}}\right)$
$\mathcal{H}_d^2$	$\Theta(d^{-M})$	$\mathcal{O}(d^M)$	$\mathcal{O}(d^M)$	$\mathcal{O}(d^{2M})$

TABLE 1. Comparison of sampling complexities for approximating f when $\mathbf{a}$ is compressible (function classes $\mathcal{F}_d^1$, $\mathcal{F}_d^2$) with those when no compressibility assumption is made on $\mathbf{a}$ (function classes $\mathcal{H}_d^1$, $\mathcal{H}_d^2$).

comparison of sampling complexities between [12] and our work for the function class $\mathcal{G}_{d,k}$.

$g_0 \in \mathcal{G}_{d,k}$	Scaling of α	$m_{\mathcal{X}}$	m_Φ	$m_{\mathcal{X}} \times (m_\Phi + 1)$
Compressible $\mathbf{A}$	$\Theta(d^{-M})$	$\mathcal{O}(kd^M \log k)$	$\mathcal{O}\left(k^{\frac{2}{2-q}} d^{\frac{Mq}{2-q}}\right)$	$\mathcal{O}\left(k^{\frac{4-q}{2-q}} d^{\frac{2M}{2-q}} \log k\right)$
Arbitrary $\mathbf{A}$	$\Theta(d^{-M})$	$\mathcal{O}(kd^M \log k)$	$\mathcal{O}(k^2 d^M \log k)$	$\mathcal{O}(k^3 d^{2M} (\log k)^2)$

TABLE 2. Comparison of sampling complexities for approximating radial functions: $f(\mathbf{x}) = g_0(\|\mathbf{A}\mathbf{x}\|_{l_2^k})$.

Remark 3. *Note that in function class denoted by $\mathcal{G}_{d,k}$, we require $g_0'(0) = 0$, since otherwise $g(\cdot)$ would not be differentiable at the origin.*

We now qualitatively demonstrate our generalization of the above function class via our Proposition 2 and highlight its significance. Assume that $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ where g has the following form:

$$g(y_1, \ldots, y_k) = \sum_{l=1}^{k} g_l(y_l). \tag{5.2}$$

We have $\dfrac{\partial g}{\partial y_i} = g_i'(y_i)$ and, $\nabla^2 g(\mathbf{y}) = \operatorname{diag}(g_1''(y_1), \ldots, g_k''(y_k))$. Clearly, $\nabla^2 g(\mathbf{0})$ is full rank if and only if $g_i''(0) \neq 0 \ \forall \ i = 1, \ldots, k$. Hence, we conclude that if the individual g_i's in (5.2) are such that for each $i = 1, \ldots, k$, we have $g_i''(0) \neq 0$, and g_i'' is Lipschitz continuous in an open neighborhood of the origin, then the function g would satisfy the conditions of Proposition 2 resulting in $\alpha = \Theta(1/d)$ for large d. To give a few practical examples of such g_i's one could think of smooth kernel functions such as Gaussian and Epanechnikov, kernels used commonly in non-parametric estimation [21]. Furthermore, the sample complexity for learning functions belonging to the class specified by Proposition 2 can be seen from Table 2 by setting $M = 1$ (since $\alpha = \Theta(1/d)$). Thus the sample complexity for abitrary $\mathbf{A}$ is $\mathcal{O}(k^3 d^2 (\log k)^2)$, while for compressible $\mathbf{A}$ it is $\mathcal{O}\left(k^{\frac{4-q}{2-q}} d^{\frac{2}{2-q}} \log k\right)$.

Remark 4. *One can think of extending the conditions of Proposition 2 so that the first M order partial derivatives are 0. However, we choose to restrict our analysis to $\mathcal{C}^2$ smooth ridge functions obeying the variation conditions as defined in Proposition 2 as it enables us to state conditions on the Hessian of g evaluated at the origin which is more intuitive to interpret and easy to verify.*

6. Numerical Experiments

We present simulation results for functions of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ with $\mathbf{A}$ being the linear parameter matrix. We assume $\mathbf{A}$ to be row orthonormal and concern ourselves only with the recovery of $\mathbf{A}$ upto an orthonormal transformation.

6.1. **Logistic function ($k = 1$).** We first take $k = 1$ and consider $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$ where g is the logistic function:

$$g(y) = \frac{1}{1+e^{-y}}.$$

One can easily verify that $C_2 = \sup_{|\beta|\leq 2} \left|g^{(\beta)}(y)\right| = 1$. Furthermore we compute the value of α through the following approximation, which holds for large d:

$$\alpha = \int \left|g'(\mathbf{a}^T\mathbf{x})\right|^2 d\mu_{\mathbb{S}^{d-1}} \approx \left|g'(0)\right|^2 = (1/16).$$

We require $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ to be greater then 0.99. We fix values of $\kappa < \sqrt{2}-1$, $\rho \in (0,1)$ and $\epsilon = 10^{-3}$. The value of $m_{\mathcal{X}}$ (number of points sampled on $\mathbb{S}^{d-1}$) is fixed at 20 and we vary d over the range 200-3000. For each value of d, we increase m_Φ till $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ reaches the specified performance criteria. We remark that for each value of d and m_Φ, we choose ϵ to satisfy the bound in Lemma 2 for the specified performance criteria given by η.

Figure 1 depicts the scaling of m_Φ with the dimension d. The results are obtained by selecting $\mathbf{a}$ uniformly at random on $\mathbb{S}^{d-1}$ and averaging the value of $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ over 10 independent trials. We observe that for large values of d, the minimum number of directional derivatives needed to achieve the performance bound on $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ scales approximately linearly with d, with a scaling factor of around 1.45.

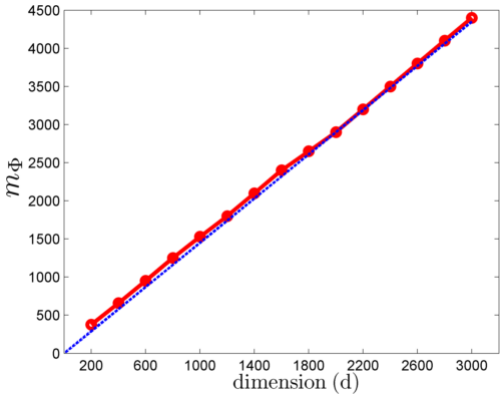

Figure 1. Plot of $\frac{m_\Phi}{d}$ versus d for $m_{\mathcal{X}} = 20$, with m_Φ chosen to be minimum value needed to achieve $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle| \geq 0.99$. ϵ is fixed at 10^{-3}. m_Φ scales approximately linearly with d with a scaling factor around 1.45.

6.2. **Sum of Gaussian functions ($k > 1$).** We next consider functions of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b}) = \sum_{i=1}^{k} g_i(a_i^T \mathbf{x} + b_i)$, where:

$$g_i(y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y+b_i)^2}{2\sigma_i^2}\right)$$

We fix $d = 100$, $\epsilon = 10^{-3}$, $m_{\mathcal{X}} = 100$ and vary k from 8 to 32 in steps of 4. For each value of k we are interested in the minimum value of m_Φ needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$. In Figure 2 we see that m_Φ scales approximately linearly with the number of gaussian atoms, k. The results are averaged over 10 trials. In each trial, we select the rows of $\mathbf{A}$ over the left Haar measure on $\mathbb{S}^{d-1}$, and the parameter $\mathbf{b}$ uniformly at random on $\mathbb{S}^{k-1}$ scaled by a factor 0.2. Furthermore we generate the standard deviations of the individual Gaussian functions uniformly over the range [0.1 0.5].

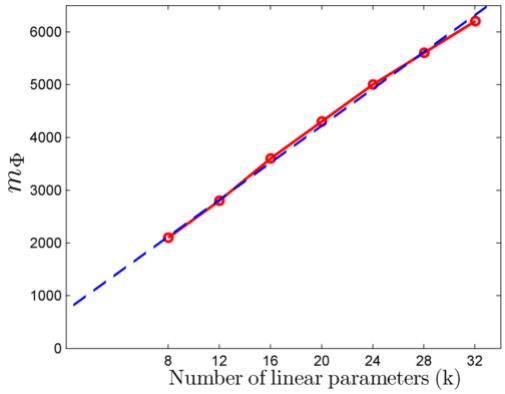

FIGURE 2. Plot of m_Φ versus k for $d = 100, m_{\mathcal{X}} = 100$, with m_Φ chosen to be minimum value needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$.

6.3. **Impact of Noise.** We now consider quadratic forms, i.e. $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x} - b\|^2$ with the point queries corrupted with Gaussian noise. Since for $g(\mathbf{y}) = \|y - \mathbf{b}\|^2$ we have $\nabla^2 g(\mathbf{b})$ to be full rank diagonal, we take α to be $1/d$. We fix $k = 5$, $m_{\mathcal{X}} = 30$, $\epsilon = 10^{-1}$ and vary d from 30 to 120 in steps of 15. For each d we perturb the point queries with Gaussian noise of standard deviation: $0.01/d^{3/2}$. This is the same as repeatedly sampling each random location approximately $d^{3/2}$ times followed by averaging. We then compute the minimum value of m_Φ needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$. We average the results over 10 trials, and in each trial, we select the rows of $\mathbf{A}$ over the left Haar measure on $\mathbb{S}^{d-1}$. The parameter $\mathbf{b}$ is chosen uniformly at random on $\mathbb{S}^{k-1}$. In Figure 3 we see that m_Φ scales approximately linearly with d.

We next repeat the above experiment under a different noise model. We are now interested in examining the scenario where a *sparse* number of point queries are corrupted with Gaussian noise. To handle this, we change the sampling scheme to random subset selection so that the i^{th} measurement takes the form: $y_i = \frac{f(\xi_j + \epsilon\phi_{i,j}) - f(\xi_j)}{\epsilon}$. This particular formulation allows us to analyse the impact of

corruption of a sparse number of queries with Gaussian noise, along the directions specified by ϕ. We use the sparCS algorithm with non convex constraints [36] for the recovery of the low rank matrix X (defined in Section 3.2). We choose the parameters $d, m_{\mathcal{X}}$ and k identically as in the previous experiment. Additionally we choose the sparsity parameter to be 1% of the number of measurements m_Φ, i.e. for each value of m_Φ, 1% of the measurements are corrupted with Gaussian noise. The standard deviation of the noise, σ is set to 0.01 as previously. By varying d from 30 to 120 in steps of 15, we compute the minimum number of measurements m_Φ needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.95$. We observe that for each d, we require to sample around 90% of the entries of the matrix X to achieve the desired approximation performance. Figure 4 shows that m_Φ scales approximately linearly with the dimension d.

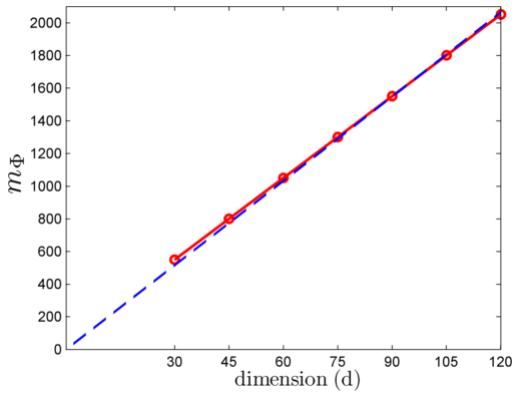

FIGURE 3. Plot of m_Φ versus d for $k = 5, m_{\mathcal{X}} = 30$, with m_Φ chosen to be minimum value needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$. Each point query is corrupted with Gaussian noise of standard deviation: $0.01/d^{3/2}$.

7. Conclusions

In this work, we consider the problem of learning multi-ridge functions of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, for arbitrary $\mathbf{A} \in \mathbb{R}^{k\times d}$ where $\text{rank}(\mathbf{A}) = k$. As compared to [12] we make no compressibility assumption on the rows of $\mathbf{A}$ thus generalizing their work to arbitrary $\mathbf{A}$. Assuming g to be a $\mathcal{C}^2$ function, our learning strategy leverages a generic stable low rank matrix recovery program to first recover an approximation $\widehat{\mathbf{A}}$ to $\mathbf{A}$ (up to an orthonormal transformation), and then uses $\widehat{\mathbf{A}}$ to form an approximation to f. We emphasize that our theoretical learning guarantees are algorithm independent as long as the low rank recovery algorithm is stable. We then establish the sampling complexity of our approach to be polynomial in the dimension d. We also provide local conditions that enable us to capture basis free sparse additive models within our framework.

Interesting future directions would involve sampling schemes for $\mathcal{C}^r$ functions with $0 < r < 2$, thus removing the current requirement that the ridge function to be approximated belong to the $\mathcal{C}^2$ class. Moreover, studying the minimax sampling lowerbounds for our approximation problem is also important. Finally, we hope to tie our analysis with the regression setting.

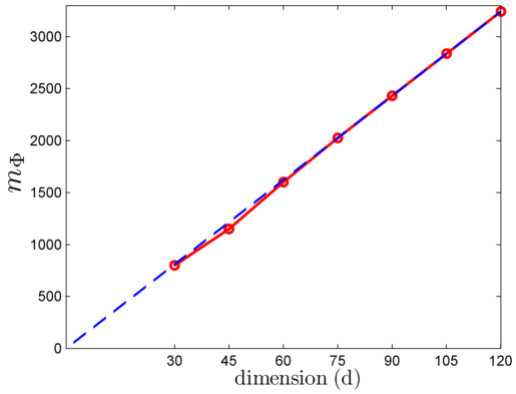

FIGURE 4. Plot of m_Φ versus d for $k = 5, m_\mathcal{X} = 30$, with m_Φ chosen to be minimum value needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$. With probability 0.01, each point query is corrupted with Gaussian noise of standard deviation: 0.01.

ACKNOWLEDGEMENTS

This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633. VC also would like to acknowledge Rice University for his Faculty Fellowship. The authors thank Jan Vybiral for useful discussions and Anastasios Kyrillidis for helping with simulations.

REFERENCES

[1] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995.
[2] E.J Candès. Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.*, 6(2):197–218, 1999.
[3] E.J Candès. Ridgelets: Estimating with ridge functions. *Ann. Stat.*, 31(5):1561–1599, 2003.
[4] E.J Candès and D.L. Donoho. Ridgelets: a key to higher dimensional intermittency? *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 357(1760):2495–2509, 1999.
[5] E.J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 2009.
[6] E.J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *CoRR*, abs/1001.0339, 2010.
[7] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
[8] E.J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56:2053–2080, May 2010.
[9] A. Cohen, I. Daubechies, R. A. DeVore, G. Kerkyacharian, and D. Picard. Capturing ridge functions in high dimensions from point queries. *Constr. Approx.*, pages 1–19, 2011.
[10] R. DeVore and G.G. Lorentz. Constructive approximation. 1993.
[11] D.L. Donoho and I.M. Johnstone. Projection based regression and a duality with kernel methods. *Ann. Statist.*, 17:58–106, 1989.
[12] M. Fornasier, K. Schnass, and J. Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012.
[13] J.H. Friedman and W. Stuetzel. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76:817–823, 1981.
[14] P. Hall and K.C. Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, pages 867–889, 1993.

[15] W. Hardle. *Applied nonparametric regression*, volume 26. Cambridge Univ Press, 1990.
[16] P.J. Huber. Projection pursuit. *Ann. Statist.*, 13:435–475, 1985.
[17] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.
[18] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.
[19] K. Lee and Y. Bresler. Admira: Atomic decomposition for minimum rank approximation. *Information Theory, IEEE Transactions on*, 56(9):4402–4416, 2010.
[20] K.C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, pages 316–327, 1991.
[21] Q. Li and J. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton, NJ, 2007.
[22] Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.
[23] B.F. Logan and L.A. Shepp. Optimal reconstruction of a function from its projections. *Duke Math. J.*, 42:645–659, 1975.
[24] L. Meier, S. Van De Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.
[25] R. Meka, P. Jain, and I.S. Dhillon. Guaranteed rank minimization via singular value projection. *CoRR*, abs/0909.5457, 2009.
[26] Th. Muller-Gronbach and K. Ritter. Minimal errors for strong and weak approximation of stochastic differential equations. *Monte Carlo and Quasi-Monte Carlo Methods*, pages 53–82, 2008.
[27] E. Novak and H. Woniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *J. Complex.*, 25:398–404, August 2009.
[28] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.
[29] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Technical Report*, 2010.
[30] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.
[31] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM REVIEW*, 52:471–501, 2010.
[32] J. Rohn. A handbook of results on interval linear problems. *Techincal Report, Czech Academy of Sciences, Prague, Czech Republic*, 2005.
[33] W. Rudin. Function theory in the unit ball of $\mathbb{C}^n$. *Springer Verlag, New York - Berlin*, 1980.
[34] J.F. Traub, G.W Wasilkowski, and H. Wozniakowski. *Information-Based Complexity*. Academic Press, New York, 1988.
[35] J. A. Tropp. User friendly tail bounds for matrix martingales. *ArXiv e-prints*, 2010.
[36] Andrew E. Waters, Aswin C. Sankaranarayanan, and Richard G. Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Neural Information Processing Systems (NIPS)*, 2011.
[37] P.A. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.
[38] H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.
[39] W.So. Rank one perturbation and its application to the laplacian spectrum of a graph. *Linear and Multilinear Algebra*, 46:193–198, 1999.
[40] Y. Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.
[41] Y. Xia, H. Tong, WK Li, and L.X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

Appendix A. Low-rank recovery formulations

We consider three distinct low-rank recovery problem settings, depending on Φ and E:

1. Affine rank minimization (ARM):. The ARM problem is exactly (3.6), where Φ is a general linear operator and the i-th entry of $\mathbf{y}$ is obtained via $[\Phi(\mathbf{X})]_i = \ll \Phi_i, \mathbf{X} \gg$. Over the last decade, several convex and non-convex algorithms address the perturbed ARM problem, such as nuclear norm minimization, matrix Dantzig selector, singular value thresholding, and ADMIRA [31, 25, 7, 8, 19]

2. Matrix completion (MC):. The MC problem revolves around a modification of (3.6) as follows

$$\mathbf{y} = \Phi_\Omega \left(\mathbf{X} + E(\mathcal{X}, \epsilon, \mathbf{\Phi}_\Omega) \right), \tag{A.1}$$

where Φ_Ω is a subset selection operator that samples a set of entries $\mathbf{X}_{i,j}, (i,j) \in \Omega$, ($|\Omega| = m_\Phi$) within the complete set of entries $[d] \times [m_\mathcal{X}]$. The ARM algorithms also handle the MC problem.

3. Robust principal component analysis (RPCA):. The original RPCA problem assumes that Φ is the identity operator so that we observe all the entries of $\mathbf{X}$. Recent generalizations also address the ARM and MC sampling formulations. An important difference compared to ARM and MC models, however, is that the RPCA approach explicitly handles unbounded outliers in observations (i.e., $\mathbf{s}(\pi)$ in (3.5)).[1] We highlight two RPCA algorithms, which relies on a convex formulation [5], and sparCS, which explicitly carries non-convex rank and sparsity constraints [36].

Appendix B. Proof of Proposition 1

Proof. By definition:

$$\|\varepsilon\|_{l_2^{m_\Phi}}^2 = \frac{\epsilon^2}{4} \left(\sum_{i=1} m_\Phi \left| \sum_{j=1}^{m_\mathcal{X}} \phi_{i,j}^T \nabla^2 f(\zeta_{i,j}) \phi_{i,j} \right|^2 \right).$$

Then, the following holds true:

$$\begin{aligned} \left| \phi_{i,j}^T \nabla^2 f(\zeta_{i,j}) \phi_{i,j} \right| &= \left| \phi_{i,j}^T A^T \nabla^2 g(\mathbf{A}\zeta_{i,j}) \mathbf{A} \phi_{i,j} \right| \\ &\leq \left\| \nabla^2 g(\mathbf{A}\zeta_{i,j}) \right\|_F \left\| \mathbf{A}\phi_{i,j} \right\|_{l_2^k}^2 \leq \frac{k^2 C_2 d}{m_\Phi}. \end{aligned}$$

Therefore,

$$\|\varepsilon\|_{l_2^{m_\Phi}}^2 \leq \frac{\epsilon^2}{4} \left(\sum_{i=1}^{m_\Phi} \left(\frac{m_\mathcal{X} k^2 C_2 d}{m_\Phi} \right)^2 \right) = \frac{\epsilon^2}{4} \frac{m_\mathcal{X}^2 k^4 C_2^2 d^2}{m_\Phi}.$$

□

[1] Here, we constrain the RPCA formulation to only the case where Φ is a subset selection operator as in (A.1).

Appendix C. Proof of Lemma 1

Proof. Let $E = \Phi^*(\varepsilon)$. We have $\|\Phi^*(\varepsilon)\| = \sup_{v,w\in\mathbb{S}^{m_{\mathcal{X}}-1}} |\langle v, Ew\rangle|$.

$$\begin{aligned}\langle v, Ew\rangle &= \mathrm{Tr}(v^T Ew) \;=\; \mathrm{Tr}(Ewv^T)\\ &= \mathrm{Tr}(\Phi^*(\varepsilon)wv^T) \;=\; \left\langle vw^T, \Phi^*(\varepsilon)\right\rangle\\ &= \left\langle \Phi(vw^T), \varepsilon\right\rangle \leq \|\varepsilon\|_{l_2^{m_\Phi}} \left\|\Phi(vw^T)\right\|_{l_2^{m_\Phi}}.\end{aligned}$$

Using Proposition 1 and since $\left\|\Phi(vw^T)\right\|^2_{l_2^{m_\Phi}} \leq (1+\kappa_1)$ holds with probability at least $1-2e^{-m_\Phi q(\kappa_1)+(d+m_{\mathcal{X}}+1)u(\kappa_1)}$, we arrive at the stated bound on $\|\Phi^*(\varepsilon)\|$. □

Appendix D. Proof of Corollary 1

Proof. Lemma 1 in conjunction with Theorem 1 gives us the following bound on $\left\|\mathbf{X}-\widehat{\mathbf{X}}_{DS}\right\|_F^2$:

$$\left\|\mathbf{X}-\widehat{\mathbf{X}}_{DS}\right\|_F^2 \leq \frac{C_0 C_2^2 k^5 \epsilon^2 d^2 m_{\mathcal{X}}^2}{4m_\Phi}(1+\kappa). \tag{D.1}$$

In general, we can have $\mathrm{rank}(\widehat{\mathbf{X}}_{DS}) > k$, thus we consider the best rank k approximation to $\widehat{\mathbf{X}}_{DS}$, in the sense of $\|\cdot\|_F$. We then obtain the following error bound:

$$\left\|\mathbf{X}-\widehat{\mathbf{X}}^{(k)}_{DS}\right\|_F \leq \left\|\mathbf{X}-\widehat{\mathbf{X}}_{DS}\right\|_F + \left\|\widehat{\mathbf{X}}_{DS}-\widehat{\mathbf{X}}^{(k)}_{DS}\right\|_F \;\leq 2\left\|\mathbf{X}-\widehat{\mathbf{X}}_{DS}\right\|_F.$$

Here, $\left\|\widehat{\mathbf{X}}_{DS}-\widehat{\mathbf{X}}^{(k)}_{DS}\right\|_F \leq \left\|\mathbf{X}-\widehat{\mathbf{X}}_{DS}\right\|_F$ as $\widehat{\mathbf{X}}^{(k)}_{DS}$ is the best rank k approximation to $\widehat{\mathbf{X}}_{DS}$ in the sense of $\|\cdot\|_F$. Finally using (D.1) we arrive at the stated bound. □

Appendix E. Proof of Lemma 2

Before beginning the proof of Lemma 2 we first recall the following theorem by [35], which provides bounds on the deviation behaviour of the largest and smallest eigenvalues of the sum of independent positive semidefinite random matrices.

Proposition 3. *(Matrix Chernoff) Consider $\mathbf{X}_1,\ldots,\mathbf{X}_m$ independent positive semidefinite random matrices of dimensions $k\times k$. Assume that $\lambda_1(X_j)\leq C$, where $\lambda_1(X_j)\geq\cdots\geq\lambda_k(X_j)$ represent the eigenvalues of X_j. Denote the eigenvalues of the sum of the expectations as*

$$\lambda_{\max} = \lambda_1\left(\sum_{j=1}^m \mathbb{E}[X_j]\right) \quad \textit{and} \quad \lambda_{\min} = \lambda_k\left(\sum_{j=1}^m \mathbb{E}[X_j]\right).$$

Then, we have the following so-called user-friendly bounds

$$\mathbb{P}\left(\left\{\lambda_k\left(\sum_{j=1}^m X_j\right)\leq(1-\rho)\lambda_{\min}\right\}\right) \leq k\exp\left(-\frac{\lambda_{\min}\rho^2}{2C}\right), \forall\rho\in(0,1),$$

$$\mathbb{P}\left(\left\{\lambda_1\left(\sum_{j=1}^m X_j\right)\geq(1+\rho)\lambda_{\max}\right\}\right) \leq k\left(\frac{1+\rho}{e}\right)^{\frac{-\lambda_{\max}(1+\rho)}{C}}, \forall\rho\in((e-1),\infty).$$

We now provide the proof of Lemma 2 below.

Proof. Observe that by Weyls inequality [38] we have $|\widehat{\sigma}_l - \sigma_l| < \tau$. Assuming $\tau < \sigma_k$ we have

$$\min_l\{\sigma_l, \widehat{\sigma}_l\} \geq (\sigma_k - \tau).$$

Thus by applying Wedins perturbation bound [37] we obtain the following bound on $\left\|\mathbf{A}^T\mathbf{A} - \widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F$:

$$\left\|\mathbf{A}_1^T\mathbf{A}_1 - \widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F = \left\|\mathbf{A}^T\mathbf{A} - \widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F \leq \frac{2}{(\sigma_k - \tau)}\left\|\mathbf{X} - \widehat{\mathbf{X}}_{DS}^{(k)}\right\|_F \leq \frac{2\tau}{(\sigma_k - \tau)}.$$

We also have the following simplified expression for $\left\|\mathbf{A}^T\mathbf{A} - \widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F$:

$$\left\|\mathbf{A}^T\mathbf{A} - \widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F^2 = 2k - 2\mathrm{Tr}(\mathbf{A}^T\mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}) = 2k - 2\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F^2.$$

This leads to the following lower bound on $\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F$:

$$2k - 2\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F^2 \leq \frac{4\tau^2}{(\sigma_k - \tau)^2} \Leftrightarrow \left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F \geq \left(k - \frac{2\tau^2}{(\sigma_k - \tau)^2}\right)^{1/2}. \tag{E.1}$$

For a non-trivial bound on $\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F$, we require the following to hold true:

$$k - \frac{2\tau^2}{(\sigma_k - \tau)^2} > 0 \Leftrightarrow \sqrt{\frac{k}{2}} > \frac{\tau}{(\sigma_k - \tau)} \Leftrightarrow \tau < \frac{\sigma_k\sqrt{\frac{k}{2}}}{(1 + \sqrt{\frac{k}{2}})}. \tag{E.2}$$

Applying Proposition 3 on (4.2) and observing that $C = kC_2^2$, we have with probability at least $1 - k\exp\left(-\frac{m_{\mathcal{X}}\alpha\rho^2}{2kC_2^2}\right)$ that $\lambda_k\left(\sum_{j=1}^m X_j\right) \geq (1-\rho)m_{\mathcal{X}}\alpha$ or equivalently $\sigma_k \geq \sqrt{(1-\rho)m_{\mathcal{X}}\alpha}$ holds true. Thus conditioning on the above event, we see that (E.2) is ensured if

$$\tau < \left(\frac{\sqrt{(1-\rho)m_{\mathcal{X}}\alpha k}}{\sqrt{k} + \sqrt{2}}\right). \tag{E.3}$$

Also, plugging the above bound on σ_k in (E.1) we obtain the stated bound on $\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F$. Lastly, observe that (E.3) is ensured if

$$\epsilon < \frac{1}{C_2k^2d(\sqrt{k} + \sqrt{2})}\left(\frac{(1-\rho)m_\Phi\alpha}{(1+\kappa)C_0m_{\mathcal{X}}}\right)^{1/2}.$$

□

Appendix F. Proof of Theorem 2

Proof. We first observe that: $\widehat{f}(\mathbf{x}) = f(\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\mathbf{x}) = g(\mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\mathbf{x})$.

$$\begin{aligned}\therefore \left|f(\mathbf{x}) - \widehat{f}(\mathbf{x})\right| = \left|g(\mathbf{A}\mathbf{x}) - g(\mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\mathbf{x})\right| &\leq C_2\sqrt{k}\left\|(\mathbf{A} - \mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}})\mathbf{x}\right\|_{l_2^k} \\ &\leq C_2\sqrt{k}\left\|\mathbf{A} - \mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F\|\mathbf{x}\|_{l_2^d}.\end{aligned}$$

Now it is easy to verify that:

$$\left\|\mathbf{A}-\mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\right\|_F^2 = \mathrm{Tr}((\mathbf{A}^T-\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\mathbf{A}^T)(\mathbf{A}-\mathbf{A}\widehat{\mathbf{A}}^T\widehat{\mathbf{A}})) = k - \left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F^2.$$

Using Lemma 2 and the fact that $\|\mathbf{x}\|_{\ell_2^d} \leq 1+\bar{\epsilon}$, we arrive at the stated approximation bound. Finally, to establish the claim in terms of δ in Theorem 2, we work our way backwards from the approximation guarantee and obtain the stated bounds. □

Appendix G. The relation of α to the hessian of f

In our Ansatz, we define α to be a lower bound on the smallest singular value of H^f in (2.1). Therefore, α is also the smallest singular value of the following matrix:

$$H^g := \int_{\mathbb{S}^{d-1}} \nabla g(\mathbf{A}\mathbf{x})\nabla g(\mathbf{A}\mathbf{x})^T d\mu_{\mathbb{S}^{d-1}}(\mathbf{x}).$$

We now note that the uniform measure $\mu_{\mathbb{S}^{d-1}}$ on the sphere $\mathbb{S}^{d-1}$ is a rotation invariant measure. For instance, if we were to project the standard rotation invariant Gaussian measure on $\mathbb{R}^d$ onto $\mathbb{S}^{d-1}$ through: $\mathbf{x}\mapsto \mathbf{x}/\left\|\mathbf{x}\right\|\ ;\mathbf{x}\in\mathbb{R}^d/\{\mathbf{0}\}$, then the resulting measure would also be rotation invariant, whereby coinciding with $\mu_{\mathbb{S}^{d-1}}$. We also observe that if we were to project the measure $\mu_{\mathbb{S}^{d-1}}$ through any $k\times d$ matrix $\mathbf{A}$ with orthonormal rows then the resultant measure μ_k is also rotation invariant and does not depend on the choice of $\mathbf{A}$.

It is a well known fact that the push-forward measure of $\mu_{\mathbb{S}^{d-1}}$ on the unit ball $B_{\mathbb{R}^k}$ is given by

$$\mu_k = \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}(1-\|\mathbf{y}\|_{l_2^k}^2)^{\frac{d-k-2}{2}}\mathcal{L}^k.$$

A proof of the above can be found for example in Section 1.4.4 of [33] where the case $\mathbb{C}^n$ is considered, which also covers the case $\mathbb{R}^n$. Based on this argument, we now arrive at the following equivalent expression for H^g:

$$H^g := \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}\int_{B_{\mathbb{R}^k}} \nabla g(\mathbf{y})\nabla g(\mathbf{y})^T(1-\|\mathbf{y}\|_{l_2^k}^2)^{\frac{d-k-2}{2}}d\mathbf{y}.$$

If the dimension $d\to\infty$ and if k is fixed, the measure μ_k concentrates around 0 exponentially fast. That is, for an open ball $B_{\mathbb{R}_k}(\epsilon)$ for a fixed $\epsilon\in(0,1)$, we have

$$\mu_k(B_{\mathbb{R}_k}(\epsilon))\to 1, \quad \text{exponentially fast as} \quad d\to\infty.$$

This phenomenon is the classical concentration of the measure $\mu_{\mathbb{S}^{d-1}}$ for large dimension d. Informally stated, the measure $\mu_{\mathbb{S}^{d-1}}$ concentrates around the equator of $\mathbb{S}^{d-1}$ as $d\to\infty$. This in turn results in the concentration of the measure μ_k around a ball of smaller and smaller radius in $\mathbb{R}^k$. We can therefore intuitively observe that the conditioning of the matrix H^g for large d would be determined predominantly by the behavior of g in a open neighborhood around the origin.

Remark 5. *If the function f is of the form $f(\mathbf{x})=g(\mathbf{A}\mathbf{x}+\mathbf{b})$ then the expression for H^g becomes the following*

$$H^g := \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}\int_{B_{\mathbb{R}^k}} \nabla g(\mathbf{y}+\mathbf{b})\nabla g(\mathbf{y}+\mathbf{b})^T(1-\|\mathbf{y}\|_{l_2^k}^2)^{\frac{d-k-2}{2}}d\mathbf{y}.$$

Denoting $B_{\mathbb{R}_k}(\mathbf{b},\epsilon)$ to be an open neighborhood around $\mathbf{b}$ for some $0<\epsilon<1$, we see that $\mu_k(B_{\mathbb{R}_k}(\mathbf{b},\epsilon))\to 1$ as $d\to\infty$. In other words, the conditioning of the matrix

H^g would now depend on the smoothness properties of g in an open neighborhood of the point $\mathbf{b}$. Keeping this in mind, we can take $\mathbf{b}$ to be $\mathbf{0}$ without loss of generality.

Appendix H. Proof of Proposition 2

Proof. Denote $\frac{\partial g}{\partial y_i} = g'_i$ and $\frac{\partial^2 g}{\partial y_i \partial y_j} = g''_{ij}$. By writing the Taylor's series of g'_i and g'_j around $\mathbf{0}$ we obtain

$$g'_i(\mathbf{y}) = g'_i(\mathbf{0}) + \sum_{l=1}^{k} y_l g''_{il}(\zeta_i),$$
$$g'_j(\mathbf{y}) = g'_j(\mathbf{0}) + \sum_{l=1}^{k} y_l g''_{jl}(\zeta_j)$$

where ζ_i, ζ_j depend on $\mathbf{y}$. Denote $H^g_{i,j}$ as the $(i,j)^{th}$ entry of H^g. We now obtain the following expression for $H^g_{i,j}$:

$$H^g_{i,j} = h_1 + h_2 + h_3, \tag{H.1}$$

where

$$h_1 = g'_i(\mathbf{0}) g'_j(\mathbf{0}), \tag{H.2}$$

$$h_2 = \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}[g'_i(\mathbf{0}) \sum_{l_2=1}^{k} \int_{B_{\mathbb{R}^k}} y_{l_2} g''_{jl_2}(\zeta_j)(1 - \|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}} d\mathbf{y} +$$
$$g'_j(\mathbf{0}) \sum_{l_1=1}^{k} \int_{B_{\mathbb{R}^k}} y_{l_1} g''_{il_1}(\zeta_i)(1 - \|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}} d\mathbf{y}], \tag{H.3}$$

and

$$h_3 = \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})} \sum_{l_1,l_2=1}^{k} \int_{B_{\mathbb{R}^k}} y_{l_1} y_{l_2} g''_{il_1}(\zeta_i) g''_{jl_2}(\zeta_j)(1 - \|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}} d\mathbf{y}. \tag{H.4}$$

We first focus on the term h_3. For some $0 < \theta < 1$, let $\mathcal{U}_\theta = B_{\mathbb{R}^k}(\theta)$ denote an open neighborhood of the origin. Then due to concentration of measure phenomenon, $\mu_k(\mathcal{U}_\theta) \to 1$ as $d \to \infty$, typically exponentially fast. Hence for large d we have the following approximation for h_3, where the approximation error decays exponentially fast with dimension (see the end of the proof for the rates):

$$\therefore\ h_3 \approx \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})} \sum_{l_1,l_2=1}^{k} \int_{\mathcal{U}_\theta} y_{l_1} y_{l_2} g''_{il_1}(\zeta_i) g''_{jl_2}(\zeta_j)(1 - \|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}} d\mathbf{y}$$
$$= B_{d,k} \sum_{l_1,l_2=1}^{k} I_{il_1,jl_2}(d,k) \tag{H.5}$$

where $I_{il_1,jl_2}(d,k) = \int_{\mathcal{U}_\theta} y_{l_1} y_{l_2} g''_{il_1}(\zeta_i) g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}} d\mathbf{y}$ and $B_{d,k} = \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}$.

Now from the Lipschitz continuity of $\frac{\partial^2 g}{\partial y_i \partial y_j}(\mathbf{y})$ in $\mathcal{U}_\theta$ we have:

$$\left|g''_{ij}(\mathbf{y}) - g''_{ij}(\mathbf{0})\right| < \theta L; \quad i,j = 1,\dots k, \quad \forall \mathbf{y} \in \mathcal{U}_\theta \tag{H.6}$$

Using (H.6) it is easy to verify the following for $\zeta_i, \zeta_j \in \mathcal{U}_\theta$:

$$g''_{il_1}(\mathbf{0})g''_{jl_2}(\mathbf{0}) - C \leq g''_{il_1}(\zeta_i)g''_{jl_2}(\zeta_j) \leq g''_{il_1}(\mathbf{0})g''_{jl_2}(\mathbf{0}) + C, \tag{H.7}$$

where $C = L^2\theta^2 + 2C_2\theta L$. We now proceed to upper bound h_3 by first considering $I_{il_1,jl_2}(d,k)$:

$$I_{il_1,jl_2}(d,k) = \int_{\mathcal{U}_\theta : y_{l_1}y_{l_2}>0} y_{l_1}y_{l_2}g''_{il_1}(\zeta_i)g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} + \int_{\mathcal{U}_\theta : y_{l_1}y_{l_2}<0} y_{l_1}y_{l_2}g''_{il_1}(\zeta_i)g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y}$$

Using (H.7) we arrive at the following upper bound:

$$I_{il_1,jl_2}(d,k) \leq \int_{\mathcal{U}_\theta} y_{l_1}y_{l_2}g''_{il_1}(\mathbf{0})g''_{jl_2}(\mathbf{0})(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} + 2C\int_{\mathcal{U}_\theta : y_{l_1}y_{l_2}>0} y_{l_1}y_{l_2}(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y}$$

Plugging the above bound on $I_{il_1,jl_2}(d,k)$ in (H.5) we get:

$$\begin{aligned}
h_3 &\lesssim B_{d,k}\sum_{l_1,l_2=1}^{k}\int_{\mathcal{U}_\theta} y_{l_1}y_{l_2}g''_{il_1}(\mathbf{0})g''_{jl_2}(\mathbf{0})(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} + \\
&2CB_{d,k}\sum_{l_1,l_2=1}^{k}\int_{\mathcal{U}_\theta : y_{l_1}y_{l_2}>0} y_{l_1}y_{l_2}(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} \\
&\leq B_{d,k}\sum_{l=1}^{k} g''_{il}(\mathbf{0})g''_{jl}(\mathbf{0})\int_{\mathcal{U}_\theta} y_l^2(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} + \\
&2CB_{d,k}\sum_{l_1,l_2=1}^{k}\int_{\mathcal{U}_\theta}(y_{l_1}^2+y_{l_2}^2)(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} \quad \left((y_{l_1}^2+y_{l_2}^2)/2 \geq y_{l_1}y_{l_2}\right) \\
&= \left(\frac{1}{k}\sum_{l=1}^{k} g''_{il}(\mathbf{0})g''_{jl}(\mathbf{0}) + 4Ck\right)B_{d,k}\int_{\mathcal{U}_\theta}\|\mathbf{y}\|^2(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} \qquad \text{(H.8)}
\end{aligned}$$

Proceeding similarly one can obtain the following lower bound:

$$h_3 \gtrsim \left(\frac{1}{k}\sum_{l=1}^{k} g''_{il}(\mathbf{0})g''_{jl}(\mathbf{0}) - 4Ck\right)B_{d,k}\int_{\mathcal{U}_\theta}\|\mathbf{y}\|^2(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y}. \tag{H.9}$$

We now focus on the term h_2. Similar to before, we have the following approximation for h_2, where the approximation error decays exponentially fast with dimension.

$$\begin{aligned}
h_2 \approx \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}[g'_i(\mathbf{0})\sum_{l_2=1}^{k}\int_{\mathcal{U}_\theta} y_{l_2}g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y} + \\
g'_j(\mathbf{0})\sum_{l_1=1}^{k}\int_{\mathcal{U}_\theta} y_{l_1}g''_{il_1}(\zeta_i)(1-\|\mathbf{y}\|^2_{\ell_2^k})^{\frac{d-k-2}{2}}d\mathbf{y}], \qquad \text{(H.10)}
\end{aligned}$$

Now it is easily verifiable that

$$\int_{\mathcal{U}_\theta} y_{l_2} g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y} = \int_{\mathcal{U}_\theta: y_{l_2}>0} y_{l_2} g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y} + \int_{\mathcal{U}_\theta: y_{l_2}<0} y_{l_2} g''_{jl_2}(\zeta_j)(1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y}$$

$$< 2\theta L \int_{\mathcal{U}_\theta: y_{l_2}>0} y_{l_2}(1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y} \tag{H.11}$$

where (H.11) follows by making use of (H.6). Through a similar process on the second summation term in (H.10) and by using $|g'_i(\mathbf{0})|, \left|g'_j(\mathbf{0})\right| < C_2$ one obtains the following upper bound on h_2.

$$h_2 \lesssim \frac{\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}[4kC_2\theta L \int_{\mathcal{U}_\theta} \|\mathbf{y}\| (1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y}] \tag{H.12}$$

One can similarly verify the following lower bound on h_2.

$$h_2 \gtrsim \frac{-\Gamma(\frac{d}{2})}{\pi^{k/2}\Gamma(\frac{d-k}{2})}[4kC_2\theta L \int_{\mathcal{U}_\theta} \|\mathbf{y}\| (1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y}]. \tag{H.13}$$

Lastly the integral term in the above bound can be bounded from above as follows.

$$\begin{aligned}\int_{\mathcal{U}_\theta} \|\mathbf{y}\| (1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y} &= \frac{2\pi^{k/2}}{\Gamma(\frac{k}{2})} \int_0^\theta r^k (1-r^2)^{(d-k-2)/2} dr \\ &< \frac{2\pi^{k/2}}{\Gamma(\frac{k}{2})} \int_0^1 r^{k-1}(1-r^2)^{(d-k-2)/2} dr \\ &= \frac{\pi^{k/2}\Gamma(\frac{d-k}{2})}{\Gamma(\frac{d}{2})}.\end{aligned}$$

Using this in (H.12) and (H.13) we obtain:

$$-4kC_2\theta L \lesssim h_2 \lesssim 4kC_2\theta L. \tag{H.14}$$

By re-writing (H.8), (H.9), (H.14) and combining with (H.2) we obtain (H.1) in matrix form:

$$H^g \precsim \nabla g(\mathbf{0})\nabla g(\mathbf{0})^T + 4kC_2\theta L\mathbf{1}\mathbf{1}^T + 4CkC_{d,k}\mathbf{1}\mathbf{1}^T + \frac{C_{d,k}}{k}\nabla^2 g(\mathbf{0})\nabla^2 g(\mathbf{0})^T, \tag{H.15}$$

$$H^g \succsim \nabla g(\mathbf{0})\nabla g(\mathbf{0})^T - 4kC_2\theta L\mathbf{1}\mathbf{1}^T - 4CkC_{d,k}\mathbf{1}\mathbf{1}^T + \frac{C_{d,k}}{k}\nabla^2 g(\mathbf{0})\nabla^2 g(\mathbf{0})^T \tag{H.16}$$

(H.17)

where $C_{d,k} := B_{d,k}\int_{\mathcal{U}_\theta} \|\mathbf{y}\|^2 (1-\|\mathbf{y}\|_{\ell_2^k}^2)^{\frac{d-k-2}{2}} d\mathbf{y}$ and $\mathbf{1}$ is a $k\times 1$ vector of all ones.

Now, we show that $C_{d,k} = \Theta(1/d)$ as $d \to \infty$. By the change of variables: $r = \|\mathbf{y}\|$, we obtain

$$C_{d,k} = \frac{2\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(k/2\right)\Gamma\left(\frac{d-k}{2}\right)} \int_0^\theta r^{k+1}(1-r^2)^{\frac{d-k-2}{2}} dr.$$

It can be checked that:

$$\int_0^{\theta} r^{k+1}(1-r^2)^{\frac{d-k-2}{2}}dr \leq \int_0^1 r^{k+1}(1-r^2)^{\frac{d-k-2}{2}}dr = \frac{1}{2}\left[\frac{\Gamma(\frac{d-k}{2})\Gamma(\frac{k+2}{2})}{\Gamma(\frac{d+2}{2})}\right]. \tag{H.18}$$

One can also verify that:

$$\begin{aligned}&\int_{\theta}^1 r^{k+1}(1-r^2)^{\frac{d-k-2}{2}}dr \leq \int_{\theta}^1 r^{k-1}(1-r^2)^{\frac{d-k-2}{2}}dr \leq e^{-(\frac{d-k-2}{2})\theta^2}\\ &\Rightarrow \int_0^{\theta} r^{k+1}(1-r^2)^{\frac{d-k-2}{2}}dr \geq \frac{1}{2}\left[\frac{\Gamma(\frac{d-k}{2})\Gamma(\frac{k+2}{2})}{\Gamma(\frac{d+2}{2})}\right] - e^{-(\frac{d-k-2}{2})\theta^2}.\end{aligned} \tag{H.19}$$

From (H.18) and (H.19) we get the following bounds for $C_{d,k}$:

$$C_{d,k} \leq \frac{k}{d}, \quad C_{d,k} \geq \left(\frac{k}{d} - \frac{2\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(k/2\right)\Gamma\left(\frac{d-k}{2}\right)}e^{-(\frac{d-k-2}{2})\theta^2}\right).$$

In other words, $C_{d,k} = \Theta(k/d)$ as $d \to \infty$, for fixed k, θ.[2]

In (H.15), we have a summation of four terms. The first three terms are rank-1 matrices with the last two vanishing as d grows. The fourth term is a full rank matrix by assumption. In this case, denote $\mathbf{V}$ as the summation of $\mathbf{D} = \frac{C_{d,k}}{k}\nabla^2 g(\mathbf{0})\nabla^2 g(\mathbf{0})^T$ and the rank-1 matrix $\mathbf{E} = \nabla g(\mathbf{0})\nabla g(\mathbf{0})^T$: $\mathbf{V} = \mathbf{D} + \mathbf{E}$. Since both matrices are symmetric positive semidefinite, we can use the singular value interlacing theorem for rank-1 perturbations [39], which states

$$\sigma_1(\mathbf{V}) \geq \sigma_1(\mathbf{D}) \geq \sigma_2(\mathbf{V}) \geq \sigma_2(\mathbf{D}) \geq \ldots \geq \sigma_{k-1}(\mathbf{D}) \geq \sigma_k(\mathbf{V}) \geq \sigma_k(\mathbf{D}). \tag{H.20}$$

Therefore, the order of the k-th largest singular value of $\mathbf{V}$ is bounded by the $(k-1)$-th and the k-th largest singular values of $\mathbf{D}$, which scale as $C_{d,k}$. In other words, $\sigma_k(\mathbf{V}) = \Theta(1/d)$.

Moreover, using results for eigenvalue bounds for symmetric interval matrices [32], we have the following bounds on the singular values of H^g:

$$\sigma_i(\mathbf{V}) - 4Ck^2C_{d,k} - 4C_2\theta Lk^2 \leq \lambda_i(H^g) \leq \sigma_i(\mathbf{V}) + 4Ck^2C_{d,k} + 4C_2\theta Lk^2. \tag{H.21}$$

where we recall that $C = L^2\theta^2 + 2C_2\theta L$. We now consider the following scenarios:

(1) If $\nabla g(\mathbf{0}) = \mathbf{0}$, then the "$4C_2\theta Lk^2$ term" in (H.21) vanishes, leading to

$$\lambda_k(H^g) \in \left[\sigma_k(\mathbf{V}) - \frac{4Ck^3}{d}, \sigma_k(\mathbf{V}) + \frac{4Ck^3}{d}\right].$$

Hence for $\theta = O(1/k^3)$, we obtain $\lambda_i(H^g) = \Theta(1/d)$.

(2) If $\nabla g(\mathbf{0}) \neq \mathbf{0}$, we obtain

$$\lambda_k(H^g) \in \left[\sigma_k(\mathbf{V}) - 4\left(\frac{Ck^3}{d} + C_2\theta Lk^2\right), \sigma_k(\mathbf{V}) + 4\left(\frac{Ck^3}{d} + C_2\theta Lk^2\right)\right]. \tag{H.22}$$

We see from (H.22) that $\lambda_k(H^g) = \Theta(1/d)$ holds provided the Lipschitz constant L is sufficiently small. In particular, if $L = O(1/d)$, then for $\theta = O(1/k^3)$ we see that $\lambda_k(H^g) = \Theta(1/d)$ holds true.

□

[2] θ can depend on k, which is not a problem since k is fixed.

Current address, H. Tyagi: ETH Zürich,
E-mail address, H. Tyagi: htyagi@inf.ethz.ch

Current address, V. Cevher: Laboratory for Information and Inference Systems (LIONS),
Ecole Polytechnique Fédérale de Lausanne,
E-mail address, V. Cevher: volkan.cevher@epfl.ch